

AMPHORA2 User Manual

An Automated Phylogenomic Inference Pipeline for Bacterial and Archaeal Sequences.

COPYRIGHT

2011 by Martin Wu

AMPHORA2 is free software: you may redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or any later version.

AMPHORA2 is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details (<http://www.gnu.org/licenses/>).

For any other inquiries send an Email to Martin Wu: mw4yv@virginia.edu

CITATION

When publishing work that is based on the results from AMPHORA2 please cite:
Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 2012; 28(7):1033-1034.

DEPENDENCY

AMPHORA2 depends on several external programs.

1. HMMER3 (<http://hmmer.janelia.org/>). Required for marker identification, sequence alignment and trimming. Earlier versions of HMMER will not work!
2. RAxML version 7.3.0 or later (<https://github.com/stamatak/standard-RAxML/downloads>). Required for phylotyping.
3. Bioperl 1.5.2 or later (http://www.bioperl.org/wiki/Getting_BioPerl).
4. EMBOSS (<http://emboss.sourceforge.net/download/>). The 'getorf' program of the EMBOSS package is required only if you analyze DNA sequences using AMPHORA2

Make sure that these programs are installed and are in your system's executable search path. To test, in a terminal type

```
raxmlHPC -version  
raxmlHPC-PTHREADS -version  
hmmsearch -h  
hmmalign -h  
getorf -help
```

If you see version or help messages, then these programs have been correctly installed. It is important to make sure they are the correct versions.

A script named 'preinstall.pl' is also included with AMPHORA2 to check and install the dependencies automatically. You need the privilege of the system administrator to run the script. See below for instructions.

INSTALLATION

1. Download AMPHORA2
2. Unpack AMPHORA2
tar -zxvf AMPHORA2.tar.gz
3. Install dependencies if they have not been installed
cd AMPHORA2
sudo perl preinstall.pl
4. Setup AMPHORA2.
You need to set up the environment variable 'AMPHORA2_home' so the AMPHORA2 scripts know where to look for the phylogenetic marker database and the NCBI taxonomy information. Let's suppose your unpacked AMPHORA folder is at /home/foo/AMPHORA2.

If you are using a bash shell, you can add the following lines to the end of the file ~/.bashrc

```
export AMPHORA2_home=/home/foo/AMPHORA2
```

Then in the terminal, issue this command

```
source ~/.bashrc
```

If you are using a C shell, you can add the following lines to the end of the file
~/.tcshrc.

```
setenv AMPHORA2_home /home/foo/AMPHORA2
```

Then in the terminal, issue this command

```
source ~/.tcshrc
```

5. Make the AMPHORA2 scripts executable.

```
chmod +x /home/foo/AMPHORA2/Scripts/*
```

You should see five folders.

1. Marker

This folder contains a seed alignment file in Stockholm format (*.stock), an alignment mask file (*.mask), a profile HMM file (*.HMM) and a tree file in newick format (*.tre) for each marker gene.

For more information about the phylogenetic markers that are included in AMPHORA2, see the marker.list file in the Marker folder.

2. Scripts

This folder contains the scripts for marker identification, alignment, trimming and phylotyping.

3. Taxonomy

This folder contains the NCBI taxonomy database that is used by the Phylotyping.pl script for phylotyping.

4. Tree

This folder contains the bacterial and archaeal genome trees in newick format. The genome trees are RAxML maximum likelihood trees made from concatenated protein sequences.

5. TestData

This folder contains the E. coli genome assembly (ecoli.fasta) and proteome sequences (ecoli.pep) for testing AMPHORA2.

RUNNING AMPHORA2

1. Marker identification

Use MarkerScanner.pl to identify bacterial and/or archaeal marker sequences. Given a sequence file, this program will identify markers from the input sequences and generate a protein fasta file for each marker gene in your working directory. For example, rpoB.pep, rpsJ.pep. When DNA input sequences are used, this program first identifies ORFs longer than 100 bp in all six reading frames, then scans the translated peptide sequences for the phylogenetic markers.

perl MarkerScanner.pl sequence-file

Options:

- DNA: input sequences are DNA. Default: no.
- Eval: HMMER eval cutoff. Default: 1e-3
- Bacteria: input sequences are Bacterial sequences
- Archaea: input sequences are Archaeal sequences
- ReferenceDirectory: the file directory that contain the reference alignments, hmms and masks. Default: /home/foo/AMPHORA2/Marker
- Help: print help

Example:

1a. Identify phylogenetic markers from the E. coli proteome.

Perl MarkerScanner.pl -Bacteria TestData/ecoli.pep

1b. Identify phylogenetic markers from the E. coli genome assembly

perl MarkerScanner.pl -Bacteria -DNA TestData/ecoli.fasta

If AMPHORA2 has been installed correctly, at the end of the run for example 1a or 1b, you should see 31 marker protein sequences (*.pep) in your working directory.

1c. If you want to identify phylogenetic markers from metagenomic sequence reads (e.g., 454 reads) of a mixed bacterial and archaeal population,

perl MarkerScanner.pl -DNA metagenomic.fasta

However, if you know your input sequences only contain bacterial or archaeal sequences, then use the -Bacterial or -Archaeal option. This makes the AMPHORA2 run faster and the results will be more accurate.

2. Marker sequence alignment and trimming

This program will align, mask and trim the marker protein sequences. Output will be aligned/trimmed sequences. For example, rpoB.aln, rpsJ.aln and their corresponding alignment masks. The alignment masks can be used to weigh the alignment columns with the RAXML's -a option (for untrimmed alignment only).

perl MarkerAlignTrim.pl

Options:

- Trim: trim the alignment using masks embedded with the marker database.
Default: no
- Cutoff: the Zorro masking confidence cutoff value (0 - 1.0; default: 0.4);
- ReferenceDirectory: the file directory that contain the reference alignments, hmms and masks. Default: /home/foo/AMPHORA2/Marker
- Directory: the file directory where sequences to be aligned are located.
Default: current directory
- OutputFormat: output alignment format. Default: phylip. Other supported formats include: fasta, stockholm, selex, clustal
- WithReference: keep the reference sequences in the alignment. Default: no
- Help: print help

Example:

perl MarkerAlignTrim.pl -WithReference -OutputFormat phylip

If AMPHORA2 has been installed correctly, at the end of the run, you should see an alignment file (*.aln) and a mask file (*.mask) for each of the 31 ecoli marker proteins in your working directory.

It is important to know that in order to run the Phylotyping.pl script properly, the MarkerAlignTrim.pl needs to be run using '-WithReference -OutputFormat phylip' options.

3. Phylotyping

Use Phylotyping.pl to assign phylotypes for each identified marker sequences. This program will assign each identified marker sequence a phylotype using parsimony method or the evolutionary placement algorithm of RAxML. The marker sequences need to be aligned first with the reference sequences using MarkerAlignTrim.pl (see above). The alignments should be in the phylip format.

perl Phylotyping.pl

Options:

- Method: use 'maximum likelihood' (ml) or 'maximum parsimony' (mp) for phylotyping. Default: ml
- CPUs: turn on the multiple thread option and specify the number of CPUs/cores to use. Important: Make sure raxmlHPC-PTHREADs is installed. If the number specified here is larger than the number of cores that are free and available, it will actually slow down the script.
- Help: print help;

Example:

assign phylotypes using the maximum likelihood method

```
perl Phylotyping.pl -CPUs 6 > phylotype.result
```

Again, if AMPHORA2 is installed correctly, you should see something like this as the output:

Query	Marker	Superkingdom	Phylum	Class	Order	Family	Genus	Species
NP_417776-NC_000913	rplB	Bacteria(0.96)	Proteobacteria(0.96)	Gammaproteobacteria(0.96)	Enterobacteriales(0.96)	Enterobacteriaceae(0.96)	Escherichia(0.50)	Escherichia coli(0.48)
NP_417097-NC_000913	rplS	Bacteria(0.96)	Proteobacteria(0.96)	Gammaproteobacteria(0.96)	Enterobacteriales(0.96)	Enterobacteriaceae(0.96)	Escherichia(0.80)	Escherichia coli(0.80)
NP_414711-NC_000913	rpsB	Bacteria(0.96)	Proteobacteria(0.96)	Gammaproteobacteria(0.96)	Enterobacteriales(0.96)	Enterobacteriaceae(0.96)	Escherichia(0.80)	Escherichia coli(0.78)

The phylotyping results are tab-delimited. The numbers within the parentheses are the confidence scores of the assignment.

KNOWN ISSUES

Bioperl

AMPHORA2 has been tested on Bioperl 1.6.1. People have reported problems running AMPHORA2 on Bioperl 1.6.9. For example, the following error has been reported when running Phylotyping.pl:

```
----- EXCEPTION: Bio::Root::Exception -----  
MSG: parse error: expected ; or ) or ,  
  
STACK: Error::throw  
STACK: Bio::Root::Root::throw  
/usr/local/share/perl/5.14.2/Bio/Root/Root.pm:472  
STACK: Bio::TreeIO::NewickParser::parse_newick  
/usr/local/share/perl/5.14.2/Bio/TreeIO/NewickParser.pm:195  
STACK: Bio::TreeIO::newick::next_tree  
/usr/local/share/perl/5.14.2/Bio/TreeIO/newick.pm:143  
STACK: main::assign_phylotype  
/usr/local/Bioinf/AMPHORA2/Scripts/Phylotyping.pl:154  
STACK: /usr/local/Bioinf/AMPHORA2/Scripts/Phylotyping.pl:70
```

Downgrading Bioperl from 1.6.9 to 1.6.1 solves the problem.

NCBI Taxonomy

If you see the following error message when you run Phylotyping.pl, you can delete the 'name2id' file in the folder AMPHORA2_home/Taxonomy/ and run the script

again. ----- EXCEPTION: Bio::Root::Exception ----- MSG: No such file or
directory AMPHORA2_home/Taxonomy/names2id STACK: Error::throw STACK:
Bio::Root::Root::throw /lib/site_perl/5.16.3/Bio/Root/Root.pm:368 STACK:
Bio::DB::Taxonomy::flatfile::_db_connect
/lib/site_perl/5.16.3/Bio/DB/Taxonomy/flatfile.pm:463 STACK:
Bio::DB::Taxonomy::flatfile::new
/lib/site_perl/5.16.3/Bio/DB/Taxonomy/flatfile.pm:144 STACK:
Bio::DB::Taxonomy::new /lib/site_perl/5.16.3/Bio/DB/Taxonomy.pm:116 STACK:
AMPHORA2_home/Scripts/Phylotyping.pl:58